VR3D: VIEW-ROBUST 3D REPRESENTATION LEARNING FOR AERIAL-GROUND PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial-ground person re-identification is a challenging task due to cross-platform viewpoint variations, which cause severe occlusion and geometric deformation. Existing methods attempt to learn view-invariant representations exclusively within the 2D image space, where drastic viewpoint variations cause the learned features to remain coupled with viewpoint bias. To address this, we propose VR3D, a View-Robust 3D Representation Learning framework that maps images into a unified 3D coordinate space to achieve view-independent feature interaction. Specifically, we introduce View-Robust 3D Representation Interaction, which leverages 3D priors extracted from single 2D observations to lift 2D appearance features into a canonical 3D space. VR3I employs 3D Geometry-Semantic Attention to establish interactions between 2D patches and 3D voxels from corresponding body parts based on their 3D spatial locations, effectively grounding 2D semantics within a 3D framework. In addition, as the reliability of these representations varies across samples due to viewpoint changes and 3D reconstruction errors, we introduce a Reliability-Aware Fusion that estimates sample-specific reliability and adaptively aggregates the multi-source representations. Extensive experiments on three benchmark datasets (CARGO, AG-ReID.v1, and AG-ReID.v2) demonstrate that VR3D outperforms recent methods, *e.g.*, it achieves an improvement of 5.63% in Rank-1 on CARGO. Our code will be released.

1 INTRODUCTION

Person re-identification aims to retrieve persons with the same identity as a query image across different cameras and plays an important role in intelligent surveillance, public safety, and other fields (He et al., 2021; Li et al., 2023; He et al., 2024; Yuan et al., 2025; Zhou et al., 2026b). In recent years, with the use of drones, aerial-ground person re-identification has attracted increasing attention. As a challenging cross-view retrieval scenario (Zhang et al., 2023; Nguyen et al., 2023; 2024; 2025), its central problem is learning stable and discriminative identity representations under viewpoint variations. Large viewpoint differences between aerial and ground cameras cause severe occlusion and geometric deformation, making the same person exhibit different visible appearances.

To tackle these challenges, existing methods mainly follow two paradigms: viewpoint disentanglement (Zhang et al., 2024; Wang et al., 2025a; Zhang et al., 2026a) and cross-view information completion (Zhang et al., 2025; Zheng et al., 2026; Wang et al., 2026). Although these methods achieve promising results, they attempt to learn view-invariant identity representations exclusively within the 2D image space. As shown in Fig. 1 (a), a 2D image records only the visible appearance under the current viewpoint, and its representation is coupled with viewpoint bias. Therefore, it remains difficult for the model to decouple these factors and learn view-invariant identity features.

To address the difficulty of learning view-invariant features in 2D, we propose to directly achieve feature interaction and representation in a view-independent space. Visual representations in 3D space are intrinsically more robust to viewpoint changes (Shang et al., 2022), providing a unified coordinate system that is not affected by camera angles. As illustrated in Fig. 1 (b), by lifting the representation into 3D space, we mitigate the bottleneck of 2D viewpoint bias, providing more complete and consistent information to facilitate the learning of view-robust features.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

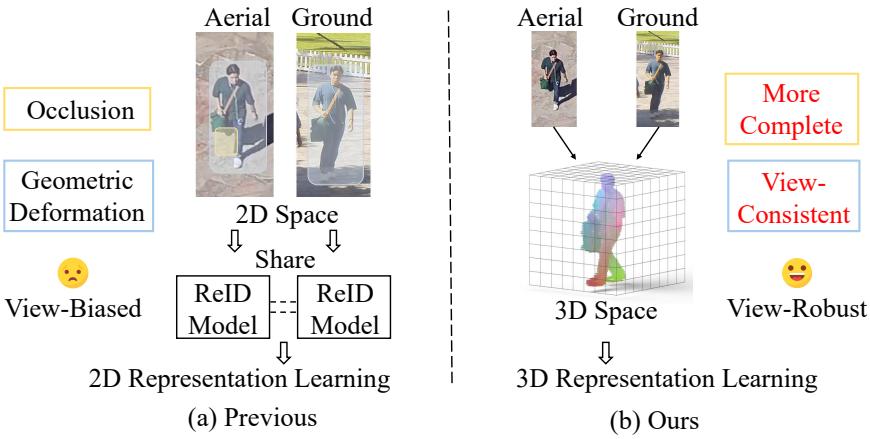


Figure 1: Illustration of our motivation. (a) Previous methods learn identity representations in 2D space, where occlusion and geometric deformation introduce strong viewpoint bias. (b) Our method learns more complete and view-consistent human representations in 3D space, enabling view-robust identity learning.

Based on this insight, we propose VR3D, a View-Robust 3D Representation Learning framework designed to achieve feature interaction and representation within a unified 3D space. To establish this 3D foundation from single 2D observations, we construct an offline pipeline that leverages the pre-trained knowledge of SAM3 (Carion et al., 2026) and SAM3D (Chen et al., 2026) to extract 3D coordinates, camera pose parameters, and voxel representations.

The 3D priors extracted by SAM3D capture geometric structures but lack the appearance semantics required for identity recognition. To address this, we propose the View-Robust 3D Representation Interaction (VR3I). By utilizing the extracted 3D coordinates to lift local 2D appearance features into a canonical 3D space, VR3I employs 3D Geometry-Semantic Attention to establish interactions between 2D patches and 3D voxels from corresponding body parts based on their 3D spatial locations. This grounds the 2D semantics within a 3D structural framework. Consequently, VR3I yields three distinct representations: an appearance feature extracted by a 2D image encoder, a view-robust 3D structural feature extracted by a 3D voxel encoder, and a spatially grounded feature that aligns appearance details based on 3D geometry.

However, the reliability of these three representations varies across samples. Extreme viewing angles reduce the reliability of 2D representations, while low-quality 3D reconstructions reduce the reliability of 3D-related representations. To prevent unreliable information from compromising identity matching, we further introduce a Reliability-Aware Fusion (RAF) that adaptively aggregates the three features into the final representation.

We evaluate VR3D on three aerial-ground person re-identification datasets, including CARGO (Zhang et al., 2024), AG-ReID.v1 (Nguyen et al., 2023) and AG-ReID.v2 (Nguyen et al., 2024). Experimental results demonstrate the promising performance of VR3D compared with recent methods, *e.g.*, it improves Rank-1 and mAP by 5.63% and 4.93%, respectively, on CARGO under the A↔G protocol. The main contributions of this paper are summarized as follows:

- We propose VR3D, a View-Robust 3D Representation Learning framework, which achieves feature interaction and representation within a unified 3D space to overcome severe aerial-ground viewpoint bias.
- We propose VR3I that leverages 3D Geometry-Semantic Attention to establish feature interactions based on physical locations. Additionally, we introduce RAF to evaluate sample-specific reliability of multi-source representations and adaptively aggregate them.
- Extensive experiments on CARGO, AG-ReID.v1, and AG-ReID.v2 demonstrate the superior performance and effectiveness of the proposed method.

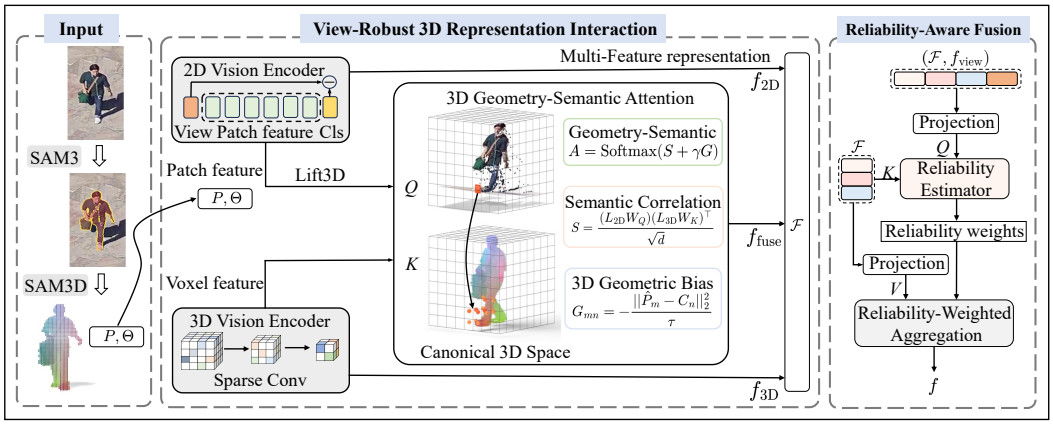


Figure 2: Illustration of the View-Robust 3D Representation Learning framework (VR3D). An offline pipeline extracts 3D human representation from the input image, followed by the View-Robust 3D Representation Interaction (VR3I) module with 3D Geometry-Semantic Attention for 2D–3D interaction in canonical 3D space to learn three complementary representations, which are adaptively aggregated by the Reliability-Aware Fusion (RAF) module based on sample-specific reliability.

2 Related Work

2.1 Aerial-Ground Person Re-Identification

Aerial-ground person re-identification aims to retrieve the same person across aerial and ground cameras, unlike traditional ReID that mainly focuses on camera networks within a single platform (Cho et al., 2022; Lee et al., 2023; Li et al., 2025; Wang et al., 2025b). Large cross-platform viewpoint differences cause substantial appearance variations, making this task highly challenging. Existing methods mainly follow two directions. The first direction learns identity representations through viewpoint-based feature disentanglement. VDT (Zhang et al., 2024), built upon ViT (Dosovitskiy et al., 2021), separates view-invariant features from view-related features. Building on this disentangled representation, DTST (Wang & Pishgar, 2025) dynamically selects identity-relevant tokens using local and global information, while SeCap (Wang et al., 2025a) introduces adaptive prompts to capture local view-invariant cues. ViSA (Zhang et al., 2026a) further models and integrates the disentangled view-invariant and view-specific features to establish cross-view semantic correspondences. The second direction introduces complementary cross-view information to compensate for incomplete identity cues from a single viewpoint. LATex (Zhang et al., 2025) and SVPR-ReID (Zheng et al., 2026) use CLIP (Radford et al., 2021) to encode textual information as cross-view semantic guidance. Other methods employ visual generative models. SD-ReID (Wang et al., 2026) uses Stable Diffusion (Rombach et al., 2022) to obtain complementary view-specific features, while 3D-LENS (Grolleau et al., 2026) employs Hunyuan3D (Zhao et al., 2025) for 3D reconstruction and novel-view rendering, generating target-view 2D images for data augmentation. These methods still alleviate viewpoint variations mainly in the 2D image space. In contrast, we investigate identity feature learning from the perspective of view-robust 3D representations.

2.2 3D Representation Learning for View Robustness

Compared with 2D appearance features, 3D representations provide more view-robust information and help alleviate appearance variations caused by viewpoint changes (Shang et al., 2022). This advantage has been exploited in several vision tasks. In robotic manipulation, 3D Diffuser Actor (Ke et al., 2025), GeoVLA (Sun et al., 2025), and Dexterity-BEV (Zhou et al., 2026a) lift 2D observations into 3D representations to improve policy adaptability to camera viewpoint changes. In pose-robust face recognition, 2D-3D Attention and Entropy (Peace et al., 2025) and NPTFace (Ran et al., 2026) leverage 3D facial geometry to learn stable identity representations. In cross-view geo-localization, Geo^2 (Zhang et al., 2026b) maps ground-level and aerial images into a shared 3D-aware space to reduce their representation gap. For conventional single-platform person re-

identification, OG-Net (Zheng et al., 2024) converts 2D person images into colored 3D point clouds and directly learns identity representations in 3D space, mitigating viewpoint changes and partial occlusion. However, these methods have not explored 3D representations for aerial-ground person re-identification. Moreover, unlike OG-Net, which relies solely on reconstructed point clouds and loses detailed appearance information during 2D-to-3D conversion, our method integrates 2D appearance semantics with 3D human representations in a canonical 3D space to obtain more complete and view-robust 3D representations.

3 METHOD

3.1 OVERVIEW

Given a query person image I and a gallery image set $\mathcal{G}$ with substantial aerial-ground viewpoint differences from I, aerial-ground person re-identification aims to retrieve person images from $\mathcal{G}$ that belong to the same identity as I. Existing methods attempt to use a shared encoder to encode images from different viewpoints. However, severe occlusion and geometric deformation introduce strong viewpoint bias into 2D observations, making it exceedingly difficult for the network to learn identity representations. To address this limitation, we propose to map images from different viewpoints into a unified 3D coordinate space to achieve direct feature interaction and representation. Based on this insight, we propose the View-Robust 3D Representation Learning (VR3D) framework, as shown in Fig. 2, to learn view-robust identity representations in this unified space.

Obtaining 3D geometric information from a single 2D image is nontrivial. Leveraging extensive 3D knowledge learned from large-scale data, SAM3D can infer 3D representations of occluded and unobserved regions from a 2D observation, thereby complementing the 3D information missing at the current viewpoint. To this end, we construct an offline 3D human representation generation pipeline that takes the original image I and the target person mask M obtained by SAM3 as input and feeds them into SAM3D. The pipeline extracts the voxel-based 3D human representation V, the 3D coordinates of the image patches P, and the camera pose parameters Θ:

$$\begin{aligned} M &= \text{SAM3}(I), \\ V, P, \Theta &= \text{SAM3D}(I, M), \end{aligned} \tag{1}$$

where $\Theta = \{R, T, s\}$, with $R \in SO(3)$ and $T \in \mathbb{R}^3$ denoting the rotation matrix and translation vector of the camera extrinsic parameters, respectively, and s is the scale factor.

Although the extracted 3D priors capture geometric information, they lack the appearance semantics required for identity recognition. To align the rich 2D appearance with the 3D geometry, we propose the View-Robust 3D Representation Interaction (VR3I). Utilizing the extracted 3D coordinates and camera pose parameters, VR3I lifts local 2D features into a canonical 3D space shared with the 3D voxels. It then employs 3D Geometry-Semantic Attention to establish interactions between 2D patches and 3D voxels from corresponding body parts based on their physical spatial locations. VR3I leads to three distinct representations: a 2D appearance feature f_{2D}, a spatially-grounded fused feature f_{fuse}, and a view-robust 3D structural feature f_{3D}:

$$\mathcal{F} = \text{VR3I}(I, V, P, \Theta), \tag{2}$$

where $\mathcal{F} = \{f_{\text{2D}}, f_{\text{fuse}}, f_{\text{3D}}\}$.

The reliability of the three representations varies across samples. Extreme viewing angles reduce the reliability of 2D appearance features, while reconstruction errors degrade 3D-related representations. To prevent unreliable information from compromising identity matching, we introduce Reliability-Aware Fusion (RAF), which uses the view-related feature f_{view} to characterize the viewing condition of each sample and adaptively aggregate the representations in $\mathcal{F}$:

$$f = \text{RAF}\left(\mathcal{F}, f_{\text{view}}\right). \tag{3}$$

3.2 VIEW-ROBUST 3D REPRESENTATION INTERACTION

To align 2D appearance with 3D geometry and bridge the cross-view appearance gap, instead of directly concatenating isolated representations, VR3I lifts local 2D appearance features into a canonical 3D space, employing 3D Geometry-Semantic Attention to directly establish feature interactions

between 2D patches and 3D voxels from corresponding human body parts based on their physical spatial locations.

We first encode the 2D image and the 3D human representation separately. For the 2D branch, we use VDT (Zhang et al., 2024) as the 2D image encoder to extract the 2D global appearance feature f_{2D}, local patch features L_{2D}, and the view-related feature f_{view}:

$$f_{\text{2D}}, L_{\text{2D}}, f_{\text{view}} = \text{2DEncoder}(I). \tag{4}$$

For the 3D branch, we employ a lightweight sparse 3D encoder to extract the local 3D geometric features L_{3D} based on the extracted voxels, followed by average pooling to obtain the view-robust 3D structural feature f_{3D}:

$$\begin{aligned} L_{\text{3D}} &= \text{3DEncoder}(V), \\ f_{\text{3D}} &= \text{AvgPool}(L_{\text{3D}}). \end{aligned} \tag{5}$$

To establish 2D-3D interaction within a unified canonical space, we utilize the patch-level 3D coordinates P and camera pose parameters $\Theta = \{R, T, s\}$ to lift the 2D image patches. The coordinate of the m-th image patch is transformed to the canonical 3D space as:

$$\hat{P}_m = \frac{(P_m - T)R^\top}{s}, \tag{6}$$

where $\hat{P}_m$ denotes the transformed coordinate of the m-th image patch, aligning the 2D appearance patches with the 3D geometric voxels.

Following this spatial alignment, we introduce 3D Geometry-Semantic Attention (3D-GSA) to establish feature interactions. 3D-GSA jointly models the geometric proximity and semantic correlation between 2D image patches and 3D voxel features. The attention weights are computed as:

$$A = \text{Softmax}(S + \gamma G), \tag{7}$$

where S denotes the semantic correlation, G denotes the 3D geometric bias, and γ controls the contribution of the geometric information.

The semantic correlation measures the content relevance between the 2D appearance features and the 3D geometric representation:

$$S = \frac{(L_{\text{2D}}W_Q)(L_{\text{3D}}W_K)^\top}{\sqrt{d}}, \tag{8}$$

where W_Q and W_K are learnable projection matrices and d denotes the feature dimension. The geometric bias provides a structural constraint based on the physical distance between each 2D image patch and 3D voxel in the canonical 3D space:

$$G_{mn} = -\frac{||\hat{P}_m - C_n||_2^2}{\tau}, \tag{9}$$

where $\hat{P}_m$ and C_n denote the coordinates of the m-th 2D image patch and the n-th 3D voxel, respectively, and τ is a distance scale factor. By jointly considering appearance semantics and spatial proximity, 3D-GSA grounds the 2D semantics within the stable 3D structural framework.

The spatially-grounded features are then computed as:

$$\begin{aligned} L_{\text{fuse}} &= L_{\text{2D}} + A(L_{\text{3D}}W_V), \\ f_{\text{fuse}} &= \text{AvgPool}(L_{\text{fuse}}), \end{aligned} \tag{10}$$

where W_V is a learnable projection matrix.

Finally, VR3I yields three distinct representations: the 2D appearance feature f_{2D}, the spatially-grounded fused feature f_{fuse}, and the view-robust 3D structural feature f_{3D}. They jointly form the output of VR3I:

$$\mathcal{F} = \{f_{\text{2D}}, f_{\text{fuse}}, f_{\text{3D}}\}. \tag{11}$$

These three representations are subsequently fed into RAF.

Table 1: Comparison with existing methods on CARGO. Results marked with † are re-evaluated by us using the released code under the official CARGO protocol. Results for SeCap are reproduced in SD-ReID (Wang et al., 2026) following the official CARGO protocol. The best performance is shown in **bold**, and the second-best performance is underlined.

Method	A↔G		ALL		G↔G		A↔A	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
VDT (Zhang et al., 2024)	48.12	42.76	64.10	55.20	82.14	71.59	82.50	66.83
DTST (Wang & Pishgar, 2025)	50.63	43.39	64.42	55.73	78.57	72.40	80.00	63.31
VIF (Khalid et al., 2025)	51.25	44.55	65.71	57.46	83.93	74.19	82.50	66.98
SeCap (Wang et al., 2025a)	48.75	46.37	64.72	56.89	82.54	75.24	82.50	66.90
SD-ReID (Wang et al., 2026)	53.12	46.44	65.06	57.47	81.25	74.08	82.50	67.70
ViSA† (Zhang et al., 2026a)	53.75	49.64	67.63	61.71	83.93	**80.01**	80.00	66.20
VR3D	**59.38**	**54.57**	**71.15**	**64.48**	**84.82**	78.74	**82.50**	**70.51**

3.3 Reliability-Aware Fusion

The reliability of $\mathcal{F}$ produced by VR3I varies across samples. The reliability of 2D features depends on the viewing condition of the current sample, while 3D-related representations are also affected by reconstruction errors reflected in their feature states. To adaptively aggregate these features, we propose a Reliability-Aware Fusion module, termed RAF, which estimates the relative reliability of each feature source by jointly considering the multi-source representations and the viewing condition of the current sample.

RAF first constructs a view-conditioned routing query by jointly encoding the multi-source identity features $\mathcal{F}$ and the view-related feature f_{view}:

$$q = \Phi_q \left(\text{Concat} \left(\mathcal{F}, f_{\text{view}} \right) \right), \tag{12}$$

where Φ_q denotes a projection layer.

To adaptively aggregate the representations, RAF formulates the dynamic fusion as cross-attention. Specifically, the multi-source representations in $\mathcal{F}$ are stacked as the keys K, while their corresponding expert projections form the values V. The attention between q and K yields sample-specific reliability weights, which are applied to V to obtain the final view-robust identity representation f:

$$f = \text{Softmax} \left(\frac{qK^\top}{\sqrt{d}} \right) V. \tag{13}$$

3.4 Training Loss

We optimize the model using identity supervision and view-related constraints. The identity loss $\mathcal{L}_{\text{id}}$ consists of the cross-entropy loss and the triplet loss, and is applied to both the 2D appearance feature f_{2D} and the final identity representation f. The view loss $\mathcal{L}_{\text{view}}$ consists of a view classification loss and an orthogonality loss that promotes view disentanglement. The overall objective is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{id}}[f_{\text{2D}}] + \mathcal{L}_{\text{id}}[f] + \lambda \mathcal{L}_{\text{view}}, \tag{14}$$

where λ balances the view-related constraint. Specifically, the identity loss is defined as

$$\mathcal{L}_{\text{id}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}}, \tag{15}$$

where $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{tri}}$ denote the cross-entropy loss and triplet loss, respectively. The view loss is defined as

$$\mathcal{L}_{\text{view}} = \mathcal{L}_{\text{vcls}} + \mathcal{L}_{\text{orth}}, \tag{16}$$

where $\mathcal{L}_{\text{vcls}}$ is the view classification loss and $\mathcal{L}_{\text{orth}}$ is the orthogonality loss.

Table 2: Performance comparison on AG-ReID.v1. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	A→G		G→A	
	R1	mAP	R1	mAP
Explain (Nguyen et al., 2023)	81.47	72.61	82.85	73.39
VDT (Zhang et al., 2024)	82.91	74.44	86.59	78.57
VIF (Khalid et al., 2025)	83.75	75.22	87.32	79.19
SeCap (Wang et al., 2025a)	84.03	76.16	87.01	78.34
SD-ReID (Wang et al., 2026)	85.16	75.40	85.97	77.02
SVPR-ReID (Zheng et al., 2026)	85.34	77.85	87.32	80.55
VR3D	**87.39**	**79.95**	**89.60**	**82.87**

Table 3: Performance comparison on AG-ReID.v2. C, W, and A denote CCTV, wearable, and aerial views, respectively. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	A→C		A→W		C→A		W→A	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Explain (Nguyen et al., 2023)	87.70	79.00	**93.67**	83.14	87.35	78.24	87.73	79.08
VDT (Zhang et al., 2024)	86.46	79.13	90.00	82.21	86.14	78.12	85.26	78.52
V2E (Nguyen et al., 2024)	88.77	80.72	93.62	84.85	87.86	78.51	88.61	80.11
SeCap (Wang et al., 2025a)	88.12	80.84	91.44	84.01	88.24	79.99	87.56	80.15
SD-ReID (Wang et al., 2026)	87.04	80.61	90.86	84.06	86.74	79.24	86.79	80.12
ViSA (Zhang et al., 2026a)	89.43	83.61	91.63	85.99	88.57	82.32	89.23	83.10
VR3D	**89.98**	**84.54**	93.57	**87.96**	**90.56**	**84.39**	**89.70**	**84.17**

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on three aerial-ground person re-identification datasets, including one synthetic dataset, CARGO (Zhang et al., 2024) and two real-world datasets, AG-ReID.v1 (Nguyen et al., 2023) and AG-ReID.v2 (Nguyen et al., 2024). CARGO contains 108,563 person images of 5,000 identities captured by 8 ground cameras and 5 aerial cameras. AG-ReID.v1 contains 21,893 images of 388 identities captured by one aerial camera and one ground camera, with aerial-view images collected at heights ranging from 15 to 45 meters. AG-ReID.v2 extends AG-ReID.v1 in identity scale and viewpoint diversity, containing 100,502 images of 1,615 identities with 15 types of attribute annotations. Its images are collected from UAVs, surveillance cameras, and smart-glass cameras. We follow the official data splits and evaluation protocols of each dataset, and use R1 accuracy and mean Average Precision (mAP) as evaluation metrics.

4.2 Implementation Details

Our method is implemented in PyTorch using the FastReID (He et al., 2023) framework and trained on a single NVIDIA A100 GPU. We use an ImageNet-pretrained Vision Transformer as the 2D backbone. All images are resized to 256×128 during training and testing. The patch size and stride are set to 16×16, with an embedding dimension of 768. The batch size is 64, containing 16 identities with 4 images per identity. The model is trained for 120 epochs using SGD. The learning rate follows a warm-up cosine decay schedule, decreasing from 8×10^{-3} to 1.6×10^{-6}. During testing, neither data augmentation nor re-ranking is applied, and the final identity representation is directly used for retrieval.

Table 4: Ablation study of key components on AG-ReID.v1."3D Global Feature" directly introduces the global 3D representation f_{3D}, "Fused Feature" denotes the fuse representation f_{fuse} learned by VR3I, and "Multi-Features" denotes average fusion of f_{2D}, f_{fuse}, and f_{3D}. RAF further performs reliability-aware fusion.

	Configuration	A→G		G→A	
		R1	mAP	R1	mAP
A	2D Baseline	82.91	74.44	86.59	78.57
B	+ 3D Global Feature	85.96	76.60	85.86	78.62
C	+ VR3I: Fused Feature	86.06	78.20	87.53	80.99
D	+ VR3I: Multi-Features	86.44	79.39	86.90	82.18
E	+ VR3I + RAF	**87.39**	**79.95**	**89.60**	**82.87**

4.3 Main Results

Table 1 reports the results on CARGO. Under the challenging A↔G cross-view setting, our method improves R1 and mAP over the previous state-of-the-art method by 5.63% and 4.93%, respectively. This result shows that our method has an advantage in aerial-ground cross-view retrieval.

Table 2 reports the results on AG-ReID.v1. Our method achieves the best performance under both the A→G and G→A protocols, reaching R1/mAP of 87.39%/79.95% and 89.60%/82.87%, respectively. Compared with the previous best method, our method improves both R1 and mAP by more than 2% under the two protocols. Compared with SD-ReID, which also uses generative priors to complement cross-view information, our method improves mAP by 4.55% under A→G and 5.85% under G→A. These results show that introducing 3D human representations effectively improves cross-view identity representation learning.

As shown in Table 3, our method achieves the best overall performance across the four protocols on AG-ReID.v2. Compared with the previous state-of-the-art method, the largest mAP improvement is 2.07% under the C→A protocol. Compared with SD-ReID, our method improves mAP by 3.93%, 3.90%, 5.15%, and 4.05% under the A→C, A→W, C→A, and W→A protocols, respectively. These results demonstrate the effectiveness of the proposed method across diverse viewpoints, including aerial cameras, fixed surveillance cameras, and wearable devices.

Overall, our method achieves performance improvements across multiple aerial-ground person re-identification datasets and evaluation protocols. These results demonstrate that our proposed method for learning visual representations in 3D space can learn more viewpoint-robust and discriminative identity features.

4.4 Ablation Study

Effects of Key Components. To evaluate the effectiveness of each component, we conduct ablation studies in Table 4. The 2D baseline uses VDT with only the 2D appearance feature f_{2D} for retrieval. Directly concatenating the 3D global feature brings only limited gains, indicating that simply introducing 3D information cannot effectively solve the viewpoint-induced appearance bias. In contrast, introducing the fused feature obtained through 2D–3D interaction in the 3D space improves mAP over the baseline by 3.76% and 2.42% under the two protocols, respectively. This result shows that our method can learn more robust cross-view identity representations. Based on the fused feature, introducing three complementary feature representations further improves the performance, which verifies their complementarity. Finally, the full model with RAF achieves the best performance and improves R1 by 2.70% over fixed average fusion under the G→A protocol, showing that reliability-aware fusion can adaptively aggregate the representations.

Effects of Key Components in 3D-GSA. To examine the roles of geometry and semantics in 3D-GSA, we compare different interaction strategies in Table 5. Using either geometric bias or semantic correlation alone outperforms coarse-grained global fusion, showing that both spatial localization and semantic correlation benefit 2D–3D interaction. Geometry performs better when used alone, indicating that physical spatial relations provide effective constraints for feature interaction. Combining both cues achieves the best results, confirming their complementarity.

Table 5: Ablation study of interaction strategies in 3D-GSA on AG-ReID.v1. "Coarse" denotes fusion by adding global 2D and 3D features, while "Geometry" and "Semantic" use only geometric bias and semantic correlation, respectively. "Geo. + Sem." combines both cues in 3D-GSA.

	Interaction	A→G		G→A	
		R1	mAP	R1	mAP
A	Coarse	86.53	78.93	86.90	81.15
B	Geometry	87.11	79.60	88.46	82.15
C	Semantic	86.82	79.48	87.84	81.74
D	Geo. + Sem.	**87.39**	**79.95**	**89.60**	**82.87**

Table 6: Ablation study of RAF on AG-ReID.v1. "Average" uses average fusion, "CA" performs cross-attention fusion, and "View" conditions fusion on view-related features.

	Configuration			A→G		G→A	
	Average	CA	View	R1	mAP	R1	mAP
A	✓			86.44	79.39	86.90	82.18
B		✓		87.01	79.54	88.46	82.15
C		✓	✓	**87.39**	**79.95**	**89.60**	**82.87**

Effects of Key Components in RAF. RAF is designed to adaptively aggregate representations with different reliability. As shown in Table 6, cross-attention outperforms fixed average fusion through adaptive aggregation of the three representations. Further incorporating the view-related feature consistently improves the results, showing that view information provides useful conditions for reliability estimation.

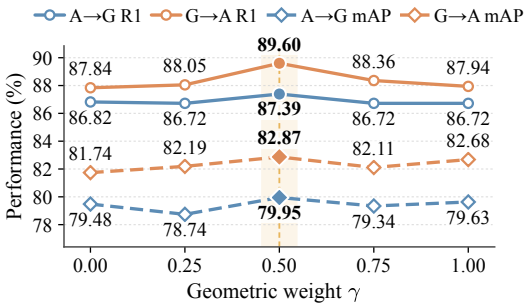


Figure 3: Effect of the geometric weight γ on AG-ReID.v1.

Hyperparameter Analysis. Fig. 3 presents the hyperparameter analysis of the geometric weight γ, which controls the spatial localization guidance in 3D-GSA. A lower γ yields insufficient geometric guidance, while a large γ overemphasizes spatial proximity at the expense of semantic relevance. Optimal performance is achieved at $\gamma = 0.50$, which is set as default.

We further analyze the retrieval results, 3D-GSA attention, and RAF routing behaviors through qualitative visualization. More details are provided in Appendix C.

5 Conclusion

In this work, we propose VR3D, a view-robust 3D representation learning framework for aerial-ground person re-identification. Unlike existing methods that learn representations in a 2D space, VR3D learns identity representations in a view-robust 3D space to reduce the effects of large aerial-ground viewpoint differences. We first construct an offline 3D human representation generation pipeline to obtain 3D information from single-person images. Based on the 3D human representation, VR3I establishes interactions between 2D semantics and 3D human representations in a canonical 3D space, producing multiple complementary identity representations. RAF then adaptively fuses these representations according to their sample-dependent reliability. Experiments on CARGO, AG-ReID.v1, and AG-ReID.v2 demonstrate the effectiveness of VR3D.

AI USE STATEMENT

During the preparation of this manuscript, we used AI tools to assist with language polishing, grammar checking, code debugging, and improving implementation efficiency. AI tools were not used to develop the core methodology, design the proposed framework, generate experimental results, or draw scientific conclusions. All AI-assisted outputs were carefully reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Xingyu Chen, FU-JEN CHU, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7220–7232, June 2026.

Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7308–7318, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

William Grolleau, Astrid Sabourin, Guillaume Lapouge, and Catherine Achard. 3d-lens: A 3d lifting-based elevated novel-view synthesis method for single-view aerial-ground re-identification. *arXiv preprint arXiv:2604.26520*, 2026.

Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9664–9667, 2023.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15013–15022, 2021.

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17521–17531, 2024.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1949–1974. PMLR, 06–09 Nov 2025.

Wajahat Khalid, Bin Liu, Xulin Li, Muhammad Waqas, and Muhammad Sher Afgan. Bridging the sky and ground: Towards view-invariant feature learning for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9749–9758, 2025.

Geon Lee, Sanghoon Lee, Dohyung Kim, Younghoon Shin, Yongsang Yoon, and Bumsub Ham. Camera-driven representation learning for unsupervised domain adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11453–11462, 2023.

Huafeng Li, Yaoxin Liu, Yafei Zhang, Jinxing Li, and Zhengtao Yu. Breaking the paired sample barrier in person re-identification: Leveraging unpaired samples for domain generalization. *IEEE Transactions on Information Forensics and Security*, 2025.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1405–1413, 2023.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2585–2590. IEEE, 2023.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024.

Huy Nguyen, Kien Nguyen, Akila Pemasiri, Feng Liu, Sridha Sridharan, and Clinton Fookes. Ag-vpreid: A challenging large-scale benchmark for aerial-ground video-based person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1241–1251, 2025.

J Brennan Peace, Shuowen Hu, and Benjamin S Riggan. 2d-3d attention and entropy for pose robust 2d facial recognition. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–11. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Zimin Ran, Xuhan Zhu, Xiang An, Xingyu Ren, Kaicheng Yang, Feilong Tang, Zhichao Chen, Yumeng Wang, Ziyong Feng, Xianzhi Wang, et al. Nptface: Native pose-aligned transformer for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1204–1213, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jinghuan Shang, Srijan Das, and Michael Ryoo. Learning viewpoint-agnostic visual representations by recovering tokens in 3d space. *Advances in Neural Information Processing Systems*, 35:31031–31044, 2022.

Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*, 2025.

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. Secap: self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22119–22128, 2025a.

Yuhai Wang and Maryam Pishgar. Dynamic token selective transformer for aerial-ground person re-identification. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29701–29710, 2025b.

Yuhao Wang, Xiang Hu, Lixin Wang, Pingping Zhang, and Huchuan Lu. Sd-reid: View-aware stable diffusion for aerial-ground person re-identification. *IEEE Transactions on Image Processing*, 2026.

Chao Yuan, Guiwei Zhang, Changxiao Ma, Tianyi Zhang, and Guanglin Niu. From poses to identity: Training-free person re-identification via feature centralization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24409–24418, 2025.

Pingping Zhang, Xiang Hu, Yuhao Wang, and Huchuan Lu. Latex: Leveraging attribute-based text knowledge for aerial-ground person re-identification. *arXiv preprint arXiv:2503.23722*, 2025.

Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22000–22009, 2024.

Quan Zhang, Zeqiang Cai, Peiming Zhao, Jingze Wu, Cailun Wu, Hongbo Chen, and Jianhuang Lai. View-aware semantic alignment for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4383–4392, 2026a.

Shizhou Zhang, Qingchun Yang, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Ground-to-aerial person search: Benchmark dataset and approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 789–799, 2023.

Yancheng Zhang, Xiaohan Zhang, Guangyu Sun, Zonglin Lyu, Safwan Wshah, and Chen Chen. Geo2: Geometry-guided cross-view geo-localization and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19432–19442, 2026b.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Aihua Zheng, Hao Xie, Xixi Wan, Zi Wang, Shihao Li, Jin Tang, and Bin Luo. Semantic-driven visual progressive refinement for aerial-ground person reid: A challenging large-scale benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13360–13368, 2026.

Zhedong Zheng, Xiaohan Wang, Nenggan Zheng, and Yi Yang. Parameter-efficient person re-identification in the 3d space. *IEEE transactions on neural networks and learning systems*, 35(6): 7534–7547, 2024.

Huayi Zhou, Wei Gao, Dekun Lu, Ruiji Liu, Zhanqi Zhang, Ziyang Zhang, Jian Chen, Wenlve Zhou, Sheng Xu, Shumin Li, et al. Dexterity-bev: Aligning 3d world and actions for generalizable robot policies learning. *arXiv preprint arXiv:2606.02274*, 2026a.

Linhan Zhou, Shuang Li, Neng Dong, Yonghang Tai, Yafei Zhang, and Huafeng Li. Hierarchical prompt learning for image-and text-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13728–13736, 2026b.

A INTRODUCTION

In this supplementary material, we provide additional implementation details and experimental results to complement the main paper. Specifically, the supplementary material is organized as follows:

1. **Implementation details**
   - Offline 3D human representation generation pipeline
   - Network details of the sparse 3D encoder, 3D-GSA, and RAF
   - Official Evaluation Protocol Description
2. **Visualization**
   - Main visualization on AG-ReID.v1.
   - Additional visualization on CARGO and AG-ReID.v2.

B IMPLEMENTATION DETAILS

B.1 3D HUMAN REPRESENTATION GENERATION

To obtain complete 3D human representations, we construct an offline 3D human representation generation pipeline consisting of two stages: target person mask generation and 3D representation reconstruction. First, SAM3 generates candidate person masks using the `person` prompt, and the target mask is selected according to its position and area. We further use the `bag` prompt to incorporate identity-related belongings into the mask, resulting in a more complete target mask.

The original image and the target mask are jointly fed into SAM3D to obtain the voxel-based 3D implicit representation, patch-level 3D coordinates, and pose parameters. To reduce computational and storage costs, we directly retain the implicit features from the decoding process as the 3D human representation.

As shown in Fig. 4, SAM3 preserves the target person and identity-related regions, while SAM3D produces complete and consistent 3D visual representations across CARGO, AG-ReID.v1, and AG-ReID.v2, providing structural cues for subsequent 2D–3D interaction. The rendered results are included only for visual inspection of the reconstruction quality.

B.2 NETWORK DETAILS

Sparse 3D Encoder. Each sample's 3D implicit representation is an 8-channel sparse voxel feature defined on a 64^3 lattice. The sparse 3D encoder contains 11 convolutional layers and progressively transforms the input into a 256-channel 8^3 sparse voxel representation. Global average pooling over the active voxels produces the 256-dimensional global 3D representation f_{3D}.

3D-GSA Settings. In 3D-GSA, the 768-dimensional 2D patch features form queries, while the 256-dimensional voxel features are projected to 768 dimensions to form keys and values. The semantic weight is set to 1.0, the geometric weight to $\gamma = 0.5$, and distance temperature to $\tau = 0.03$.

RAF Settings. RAF uses four attention heads, each with a dimension of $768/4 = 192$. The three source representations are first projected into a 768-dimensional space. The projected representations are concatenated with the 768-dimensional view-related feature f_{view} and projected to form the routing query. They are also used to construct routing keys and expert features. The reliability-weighted outputs are concatenated and projected to obtain the identity representation.

B.3 OFFICIAL EVALUATION PROTOCOL DESCRIPTION

We follow the official CARGO evaluation protocol in all experiments. Specifically, under the bidirectional cross-view retrieval setting, same-view negative samples are not removed from the gallery. For SeCap (Wang et al., 2025a), we report the results reproduced by SD-ReID (Wang et al., 2026), which follows the official protocol. However, SeCap removes same-view negative samples during data loading when forming the query-gallery pairs for evaluation, thereby reducing the interference from same-view distractors. Therefore, its results are not directly comparable with those obtained



(a) CARGO (b) AG-ReID.v1 (c) AG-ReID.v2

Figure 4: Representative 3D person representation generation results on CARGO, AG-ReID.v1, and AG-ReID.v2. From left to right: input image, target person mask, voxel-based 3D implicit representation, and rendered 3D human.



Figure 5: Retrieval results on AG-ReID.v1 under the A→G protocol and G→A protocol. Green and red boxes denote correct and incorrect matches, respectively.

under the original CARGO protocol. We adopt the official evaluation protocol and verify the implementation based on the released code.

C VISUALIZATION

C.1 MAIN VISUALIZATION ON AG-REID.V1

Retrieval Results Figure 5 presents retrieval results on AG-ReID.v1 under the A→G protocol. Compared with the baseline, our method retrieves more correct cross-view matches among the top-ranked results and reduces interference from visually similar identities, demonstrating more discriminative cross-view identity representations.

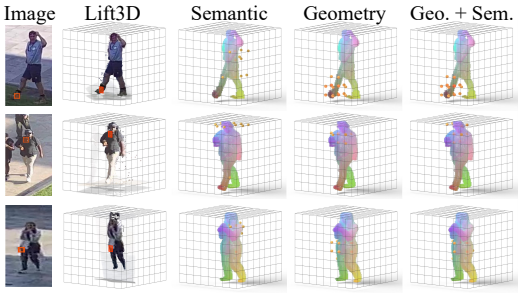


Figure 6: Visualization of 3D-GSA on AG-ReID.v1. Orange boxes indicate the query patches, while orange points denote the top-k 3D voxels selected by semantic correlation, geometric bias, and final attention, respectively.

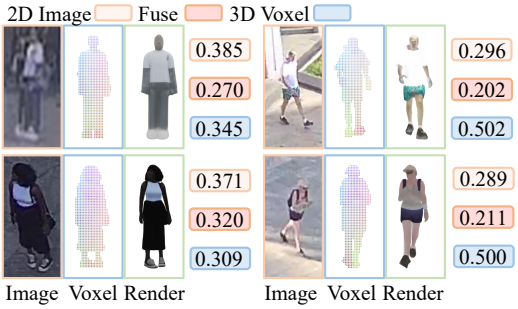


Figure 7: Visualization of RAF reliability weights under different samples on AG-ReID.v1. RAF assigns weights to f_{2D}, f_{fuse}, and f_{3D} according to their sample-dependent reliability.



(a) CARGO (b) AG-ReID.v2

Figure 8: Retrieval results of the baseline and VR3D on CARGO under the A→G protocol and on AG-ReID.v2 under the A→C protocol.

3D-GSA Attention Analysis. To examine the geometric and semantic cues in 3D-GSA, Fig. 6 visualizes the top-k voxels selected by semantic correlation, geometric bias, and final attention. Geometric bias favors spatially close voxels but lacks semantic consistency, whereas semantic correlation selects similar regions from unrelated body parts. By combining both cues, the final attention focuses on voxels that are spatially close and semantically consistent with the query patch, enabling reliable 2D–3D interaction in the canonical 3D space.

RAF Routing Analysis. To verify how RAF estimates the reliability of different feature sources across sample conditions, we visualize its routing weights in Fig. 7. RAF gives more weight to f_{2D} when 3D reconstruction is unreliable, and to f_{3D} when extreme viewing angles weaken 2D appearance but 3D reconstruction remains reliable. When both 2D appearance and 3D feature are reliable, the contribution of f_{fuse} increases. These results show that RAF adaptively aggregates the three representations according to their reliability.

C.2 Additional Visualization on CARGO and AG-ReID.v2

CARGO is a large-scale benchmark collected by multiple aerial and ground cameras, with substantial cross-view variations. AG-ReID.v2 covers UAV, CCTV, and wearable-camera views, providing a complementary evaluation across different camera platforms. To verify the effectiveness of VR3D across these datasets, we visualize its retrieval results, 3D-GSA attention, and RAF routing weights.

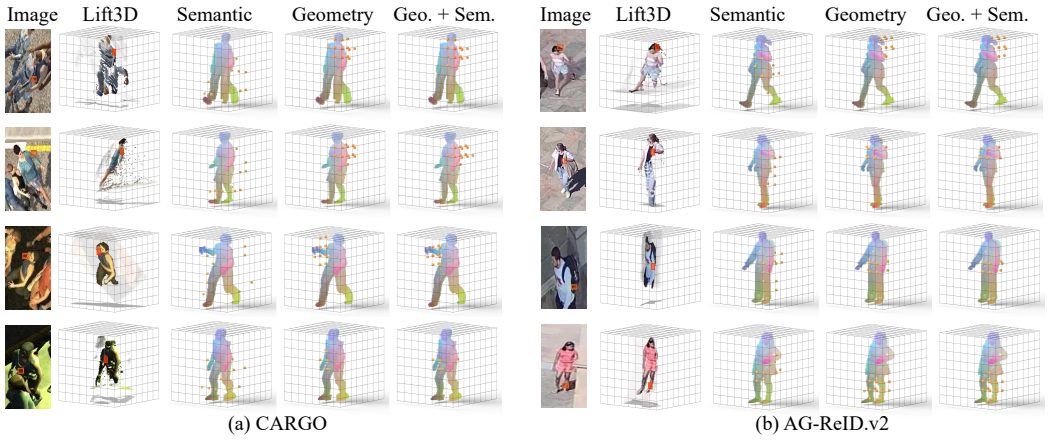


Figure 9: Visualization of 3D-GSA on CARGO and AG-ReID.v2. Orange boxes indicate the query patches, while orange points denote the top-k 3D voxels selected by semantic correlation, geometric bias, and final attention, respectively.



Figure 10: Routing weights of RAF on CARGO and AG-ReID.v2.

Retrieval Results Fig. 8 compares the rank lists of the baseline and VR3D on CARGO under the A→G protocol and on AG-ReID.v2 under the A→C protocol. This comparison examines whether the learned 3D representations improve identity matching under large viewpoint changes.

3D-GSA Visualizations We visualize the geometric bias, semantic correlation, and final attention in Fig. 9. Geometric bias favors nearby voxels, while semantic correlation identifies related but sometimes spatially inconsistent regions. The final attention concentrates on voxels supported by both cues, verifying that 3D-GSA performs spatially grounded 2D–3D interaction.

RAF Visualizations To verify that RAF performs adaptive rather than fixed fusion, Fig. 10 presents the routing weights of the three representations under sample conditions. In the first CARGO example, the inaccurate 3D reconstruction leads to a low weight for f_{3D}, while f_{2D} and f_{fuse} receive higher weights. For samples with reliable 3D structure, RAF increases the contribution of f_{3D} or assigns more balanced weights. The AG-ReID.v2 examples further show that RAF adjusts the dominant feature source according to the appearance and structural cues. In the second example, f_{fuse} receives the highest weight when both 2D and 3D information are useful.